BACKING-OFF CONTEXT- & GENDER-DEPENDENT MODELS FOR BETTER ARTICULATORY FEATURE EXTRACTION

Supphanat Kanokphara and Julie Carson-Berndsen

Department of Computer Science
University College Dublin
Ireland
Tel: (353 1) 716 2493
Fax: (353 1) 269 7262
{supphanat.kanokphara, julie.berndsen}@ucd.ie

The majority of speech recognition systems today commonly use Hidden Markov Models (HMMs) as acoustic models in systems since they can powerfully train and map a speech utterance into a sequence of units. Such systems perform even better if the units employed are context-dependent and gender-dependent. Analogously, when HMM technology is applied to the problem of articulatory feature extraction, context- and gender-dependent articulatory features should definitely yield a better result. This paper presents a possible strategy which utilizes the strength of context- and gender-dependent models to build a better HMM-based articulatory feature extraction system.

Keywords: Articulatory feature extraction, Speech recognition, Context-dependent model, Gender-dependent model, Hidden Markov Model, Articulatory feature.

Introduction

HMMs are predominantly used as acoustic models in current speech recognition systems. The reason for this is that HMMs can normalize the time-variation of the speech signal and characterize the speech signal statistically in the optimal sense. However, by completely ignoring linguistic knowledge and relying only on statistical models, such systems can achieve only a limted level of success (Lee, 2004). One of the problems of stochastic systems is that they are too restrictive and thus not fully applicable in adverse environments (noisy, out-of-task, out-of-vocabulary, etc). Many researchers are aware of this problem and as a result there have been moves to try to integrate more explicitly knowledge-based and statistical approaches. One of the ways in which such a hybrid system can be achieved is to employ articulatory features as the first level of recognition (Carson-Berndsen, 1998), (Kirchhoff, Fink and Sagerer, 2002) and (Richardson, Bilmes and Diorio, 2003), rather than phones, phonemes or higher level units. With this approach, more linguistic knowledge can be incorporated into the system and hence it yields better recognition results. In order to use articulatory features for speech recognition, many articulatory feature extraction systems have been developed (Kanokphara and Carson-Berndsen, 2004), (Abu-Amer and Carson-Berndsen, 2003), (Chang, Greenberg and Wester, 2001) and (Ali et al. 1999). Among these, the HMM-based systems appear to outperform the others. While we believe that in the longer term, feature extraction systems should be tailored to specific types of acoustic-phonetic information, in this paper we restrict our discussion to ways in which the performance of an HMM-based feature extraction system can be improved using context- and gender-dependent models.

Since an HMM-based system trains models optimally without knowing exactly what it is training, it is crucial that the models designed accurately represent and recognize speech. To do this effectively, contextual effects which usually cause degradation in the system, have to be accounted for in the model design. In this paper, two specific contextual factors are addressed: gender and context. These factors are typically used to upgrade standard (higher-level unit) HMM-based speech recognition systems. Similarly, we choose to upgrade HMM-based articulatory feature extraction systems with context- and gender-dependent features. In what follows, we examine possible ways to broaden context-independent units into context-dependent counterparts. We bring in the idea of gender-dependency. We discuss a standard HMM-based articulatory feature extraction system which does not avail of context- and gender-dependent information. Then an extension to include context- and gender-dependent features is suggested. The context- and gender-dependent system is compared with a purely context-dependent system (Kanokphara and Carson-Berndsen, 2004) and the results are presented. Finally, conclusions are drawn and future directions discussed.

Context-Dependent Units for Speech Recognition

One of the difficulties discussed most in connection with the introduction of context-dependent units into speech recognition systems is how to strike a balance between the level of information in models and the limited acoustic training data. This is because the number of context-dependent units is naturally large. To illustrate this, let "a" be a context-independent unit. Then, its context-dependent version can be labeled as "b-a+c" where "b" and "c" are preceding and succeeding units, respectively. Therefore, if N is the number of context-independent units, the number of context-dependent counterparts will be NxNxN which is unacceptable for training. In order to address this problem, the suitability of three strategies for making context-dependent units trainable were examined, namely, backing-off, smoothing and sharing. Each technique requires determination of parameters to be backed-off to, smoothed with and shared with others, respectively; they are discussed briefly here.

Backing-Off

Backing-off is the simplest strategy for training context-dependent units. When insufficient data for training a model exists, that model backs-off and some less informative but trainable model is used instead. For example, if a triphone has only a few examples in the training data, a biphone should be used. If a biphone is still not trainable, monophone should be used. With this strategy, it is possible to insure that all models are well trained. The disadvantage of this strategy, however, is that the difference between more and less informative models is too large when a backing-off occurs.

Smoothing

Lee and Hon (1989) proposed an alternative way to keep a balance between information in models and sufficiency of training data. The *smoothing* method uses interpolation between less informative but trainable and more informative but un-trainable models. The advantage of this strategy is that it can smooth deeply into the state level, in contrast to the backing-off strategy which is applied only at the model level. However, the smoothing technique has not been widely applied in speech recognition and for this reason we do not consider this technique any further here.

Sharing

The *sharing* strategy is perhaps the most frequently used for balancing trainability and information of models. Sharing schemes can be divided into two approaches, namely bottom-up and top-down approaches. The bottom-up approach starts by generating all context-dependent units which occur in the training data. Various algorithms are available to find similar states and tie them together. In this way, tied states use the same training data and make the system trainable. However, some examples are required for searching similar states. This makes defining good unseen models impossible. The top-down approach, on the other hand, uses linguistic knowledge to form a decision tree. This tree then is used to cluster and tie states hierarchically. This tree can also synthesize unseen models linguistically and therefore it does not suffer from the same problem as the bottom-up approach (Odell, 1995).

Gender-Dependent Units for Speech Recognition

Since speech signals from each person differ, ideally, for the purposes of speech recognition, it would be better to construct individual acoustic models for each speaker. Unfortunately, it is impossible to train models for every speaker. To make this practical, two possible techniques are commonly used.

Group-Dependent Modeling

Even though each speaker has an individual speaking style, there is agreement that similarities among particular groupings of speakers can be found which allow commonalities to be classified with respect to gender, age, dialect, etc. For example, male and female speech signals are more different than inter-male speech signals. By including gender-dependent models in the system i.e., separate models for male and female speakers, specific characteristics of this speaker grouping can be characterized.

Speaker Adaptation

Another way to model speaker-dependent characteristics in a speaker-independent system is to construct speaker-dependent models during recognition. This can be done by speaker-adaptation (Leggetter and Woodland, 1994). To do this, first a speaker-independent system is constructed then speaker-dependent models are adapted from the first few sentences from an arbitrary speaker. This can be done with or without providing correct sentences for adaptation data. If correct sentences are given, it is called *supervised adaptation*. If the correct sentences are not given, it is called *unsupervised adaptation*. In either system, the adaptation data is required.

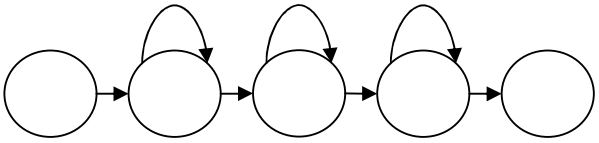

Figure 1 5-state left-right HMM

Since this paper focuses on the development of a system without adaptation, only gender-dependent models are considered

HMM-Based Articulatory Feature Extraction System

In general, an HMM is used to map some uncertain signal onto a sequence of symbolic units. These units can be words, syllables, demi-syllables, phones, etc. The articulatory feature extraction presented here also uses HMMs to map a speech signal onto a sequence of features on separate tiers (or levels). Moreover, since each HMM-based system recognizes a sequence of features on each tier independently, it allows for overlap among features on each tier thus modeling coarticulation phenomena. The feature-based approach allows for the integration of articulatory information into a statistical system and thus results in a better, more linguistic, system. In the statistical sense, the number of classes for a system to recognize is reduced and hence more robust models can be built.

System Overview

Before performing articulatory feature extraction, it is necessary to design a feature table which assigns different features to each tier. The resulting feature table (see also Kanokphara and Carson-Berndsen, 2004) contains 6 different tiers defining the articulatory space: manner, place, voicing, vowel type, vowel height and lip rounding. In this case the feature table is specifically for English data but alternative feature tables could be employed such as those suggested by (Geumann, 2004).

The articulatory feature extraction system has been constructed using HTK (http://htk.eng.cam.ac.uk/). The acoustic model training system starts by converting a speech signal into a sequence of vector parameters with a fixed 25ms frame and a frame rate of 10ms. Each parameter is then pre-emphasized with the filter $1\text{-}z^{0.97}$. The discontinuities at the frame edges are smoothed using a Hamming window. A Fast Fourier Transform is used to convert time domain frames into frequency domain spectrums. These spectrums are averaged into 24 triangular bins arranged at equal Mel-Frequency intervals (where $f_{mel} = 2595\log_{10}(1+f/700)$). 12 dimensional Mel-Frequency Cepstral Coefficients (MFCCs) are then obtained from the cosine transformation and lifter**.** The normalized log energy is also added as the 13^{th} front-end parameter. The actual acoustic energy in each frame is calculated and the maximum determined. All log energies are then normalized with respect to maximum and log energies below a silence floor (set to -50DB) clamped to that floor. The 13 front-end parameters are then expanded to 39 front-end parameters by appending first and second order differences of the static coefficients. The parameters employed here have been used extensively (Davis and Mermelstein, 1980) and have proved to be one of the best choices for HMM-based speech recognition systems.

Flat start training is then used for model initialization according to the feature table. Flat start training is a training strategy provided by HTK which requires no time-annotated training transcriptions for model initialization. Each model contains 5 states and the covariance matrices of all states are diagonal. Figure 1 shows a 5-state left-right HMM as is used in the system. Maximum likelihood estimators are used to train the HMM parameters (Juang, 1985). The number of training iterations after each change is determined automatically in line with (Tarsaku and Kanokphara, 2002). The models are finally expanded to 15 mixtures. The language model in this paper is trained from the training set on each tier using back-off bigram. The language model provides feature constraints that correspond to the inter-feature model transition probabilities. For the recognition process, the Viterbi algorithm is used without any pruning factor.

The Data

Experiments with the HMM-based articulatory feature extraction system were performed using the standard TIMIT corpus (Garofolo et al. 1993) which consists of 3600 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the U.S., of which 462 are in the training set and 168 are in the test set. There is no overlap between the training and test utterances, except 2 dialect (SA) sentences which were read by all speakers. The training set contains 4620 utterances and the testing set 1680 (112 males and 56 females). The core test set, which is the abridged version of the complete test set, consists of 192 utterances, 8 from each of 24 speakers (2 males and 1 female from each dialect region). Normally, SA sentences are eliminated from the training/test set because they occur in both training and test set. All utterances were recorded in a noise-isolated recording booth. The speech was directly digitized at a sample rate of 20 kHz with the anti-aliasing filter at 10 kHz. The speech was then digitally filtered, debiased and

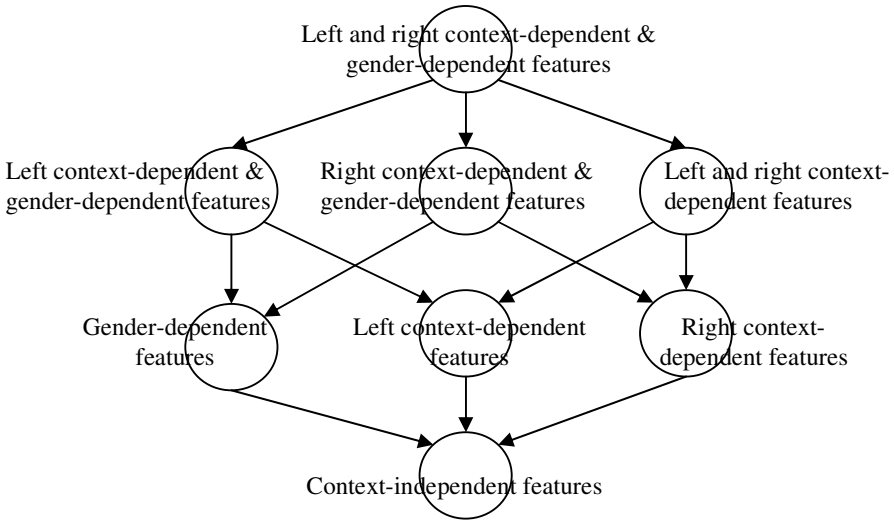

Figure 2 Backing-off hierarchy

downsampled to 16 kHz. The training/testing set used in this paper corresponds exactly to the set used in (Kanokphara and Carson-Berndsen, 2004) which is the full training set (4620 utterances) and test set without SA (1344 utterances). All TIMIT phonemic training and testing transcriptions are transformed to feature transcriptions automatically.

Context- and Gender-Dependent Articulatory Feature Extraction System

According to the discussion of the many possible techniques for making context-dependent units trainable above, the top-down approach would seem to be the best way of balancing the parameters of the model. However, this technique requires some linguistic knowledge in terms of a set of phonetic questions (a list of unit classes defined with respect to unit features). For context-dependent units in speech recognition, the units are usually phones or larger units in which there are always some mutual feature characteristics which can be shared. However, in our system, we want to construct context-dependent features which have no further mutual phonetic characteristics for sharing. The top-down approach does not appear appropriate for context-dependent features, therefore and hence other strategies must be considered. The bottom-up approach must be discarded as it is weak for unseen context-dependent units. The backing-off approach, despite its disadvantages, is chosen here for the following reasons. Firstly, the number of feature classes required by the system is obviously less than the number of phone or larger unit classes. Therefore, it is unlikely that this approach will result in too many context-independent units in the system. Secondly, it is easier to integrate gender dependency using this technique than with the smoothing technique. The result from (Lamel and Gauvain, 1993) also confirms that the backing-off technique with gender-dependent models is better than the smoothing technique. It is important to note that our technique for modeling is not the same as that described in (Lamel and Gauvain, 1993), where backed-off context dependent models are constructed first and the models then are expanded to be gender dependent. In this paper, context and gender dependencies are backed-off together as demonstrated below.

The system starts by training a single mixture context- and gender-independent system and a single mixture context-independent but gender-dependent system separately. Then context- and gender-dependent transcriptions and context-dependent but gender-independent transcriptions are expanded using a cross-word network. The training algorithm is the same as for the context-independent system except that the training transcription is changed to be backed-off to a context- and gender-dependent version. After the models are trained, the number of mixtures is then again increased to 15.

Context and gender dependency are determined automatically according to their frequencies in the training transcriptions. When left-right-context and gender-dependent feature frequency is less than 40, the highest frequency feature among left-context and gender-dependent, right-context and gender-dependent and left-right-context-dependent features is used. If all left-context and gender-dependent, right-context and gender-dependent and left-right-context-dependent features are less than 40, the highest frequency feature among left-context-dependent, right-context-dependent and gender-dependent feature is used. If all left-context-dependent, right-context-dependent and gender-dependent features are less than 40, context-independent feature is used. Figure 2 illustrates the backing-off hierarchy.

	place	manner	vowel height	vowel type	round	voice
% correct	78.48	90.54	86.08	93.62	92.51	97.12
% accuracy	73.36	81.53	80.03	68.08	86.08	72.94
No. of models	561	121	91	48	45	18

Table 1 Result from the context-dependent system from (Kanokphara and Carson-Berndsen, 2004)

	place	manner	vowel height	vowel type	round	voice
% correct	78.05	***90.583***	***86.61***	***94.54***	92.44	***97.32***
% accuracy	72.11	80.99	***80.72***	***71.09***	85.85	71.97
No. of models	931	213	179	94	83	37

Table 2 Result from context & gender-dependent system

System Performance

Percentage correct and accuracy of recognized feature sequences are used to evaluate system performance using standard techniques. Typically each of the recognized feature sequences is matched against reference label sequences by performing an optimal string alignment using dynamic programming. Once the optimal alignment has been found, the number of features, substitution features, deletion features and insertion features are counted and calculated. The difference between percentage correct and accuracy is that percentage correct ignores insertion features while percentage accuracy does not.

As vowels are more influenced by gender than consonants, our expectations were that the introduction of gender information into the system would lead to an improvement in recognition of vowel quality features. We present a comparison with the context-dependent system. Table 1 depicts the results of the context-dependent system (Kanokphara and Carson-Berndsen, 2004) and table 2 presents the results of the context- and gender-dependent system. Bold italic font in table 2 indicates better results than table 1. The results are as expected. There are improvements on vowel height and vowel type. In fact, we also expect to have an improvement on round tier as well. However, the %correct and %accuracy for round tier are only a little dropped.

It also can be observed that the number of context- and gender-dependent models is nearly double those of context-dependent only models. This means that there are only small amount of gender-independent models used in the system. The highest gender backing-off percentage in this experiment occurs on the place tier which is approximately 17%. The lowest gender backing-off percentage is on the voice tier because no gender-dependent model is backed-off. Furthermore, there is one extra model added to the system due to unbalanced numbers of male and female in the corpus. This convinced us that including gender dependency in the backing-off process is better than only context dependency backing-off. For example, on the place tier, if gender-dependency is not included in the backing-off process, the number of the total models has to be doubled from 561 to 1122. This would lead to a worse result since the number of models is too large and the information adding to models is inappropriate (gender in this case).

Conclusion

There are many aspects of this research worth emphasizing here. Firstly, to implement better speech recognition systems, hybrid approaches which use statistical and linguistic knowledge are very attractive. Applying articulatory features is one of the possible ways to integrate linguistic knowledge into stochastic speech recognition systems. In order to build good articulatory-feature-based systems, reliable articulatory feature extraction systems have to be studied and researched. In this paper, we proposed an alternative way to efficiently extract articulatory features from speech utterances. Secondly, as articulatory features are common to most languages, this makes our system language-independent, although clearly the set of features does have to be extended beyond the set used in this paper (Geumann, 2004). Thirdly, as our system is based on HMMs, many useful techniques for HMM-based speech recognition systems can also be applied to our system. However, articulatory feature extraction and phone (or larger unit) recognition systems cannot be treated in exactly the same way. Some technique has to be customized for articulatory feature training. For example, in this paper, in order to use context-dependent technique, the back-off approach was preferred

over the top-down approach. Fourthly, our system can also be extended further by integrating more linguistic knowledge into our system. For example, on the voice tier, if a segment of speech utterance is recognized to be unvoiced, it cannot be recognized to be vocalic on manner tier. Finally, even though in HMM-based speech recognition system, gender and context dependencies are proved to be useful, it seems that only context dependency is good for articulatory feature extraction systems. This is due to the fact that gender information highly dominates over vowels more than consonants. Nevertheless, context together with gender dependency still can yield some improvement on vowel quality tiers. Future work is concerned with investigating the application of other feature tables in this model.

Acknowledgements

This material is based upon works supported by the Science Foundation Ireland for the support under Grant No. 02/IN1/I100. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

References

Abu-Amer, T. and Carson-Berndsen, J. 2003. HARTFEX: A Multi-Dimensional System of HMM Based Recognizers for Articulatory Feature Extraction, In Proc. Eurospeech. Geneva, Switzerland.

Ali, A. M. A., Van der Spiegel, J., Mueller, P., Haentjaents, G. and Berman, J. 1999. An Acoustic-Phonetic Feature-Based System for Automatic Phoneme Recognition in Continuous Speech, In Proc. IEEE Int. Symposium on Circuits and Systems (ISCAS), 118-121.

Carson-Berndsen, J., 1998. Time Map Phonology: Finite State Models and Event Logics in Speech Recognition. Kluwer Academic Publisher, Dordrecht.

Chang, S; Greenberg, S. and Wester, M. 2001. An Elitist Approach to Articulatory-Acoustic Feature Classification, In Proc. Eurospeech, Aalborg.

Davis, S.B., Mermelstein, P. 1980. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. Acoustic Speech and Signal Processing* 28(4). 357-366.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L. 1993. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM, NIST.

Geumann, A. 2004. Towards a New Level of Annotation Detail of Multilingual Speech Corpora. In Proc. Int. Conf. Spoken Language Processing, Jeju Island, Korea.

HTK Speech Recognition Toolkit, http://htk.eng.cam.ac.uk/, Cambridge University, Engineering Department.

Juang, B.H. 1985. Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains. *AT&T Tech. J.* 64(6).

Kanokphara, S. and Carson-Berndsen, J. 2004. Better HMM-Based Articulatory Feature Extraction with Context-Dependent Model, submitted to The 18th International Florida Artificial Intelligence Research Society Conference.

Kirchhoff, K., Fink, A., G. and Sagerer, G. 2002. Combining Acoustic and Articulatory Feature Information for Robust Speech Recognition. *Speech Communication* 37: 303-319.

Lamel, L.F., Gauvain, J.L. 1993. High Performance Speaker-Independent Phone Recognition Using CDHMM. In proc. In Proc. EuroSpeech, 121-124. Berlin.

Lee, C. 2004. From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition. In Conference Program and Abstract Book, Eighth International Conference on Spoken Language Processing, 109-112. Jeju Island, Korea.

Lee, K.F., Hon, H.W. 1989. Speaker-Independent Phone Recognition Using Hidden Markov Models. *IEEE Trans. Acoustic Speech and Signal Processing* 37(11).

Leggetter, C.J., Woodland, P.C. 1994. Speaker Adaptation of Continuous Density HMMs using Multi-variate Linear Regression. Proc. Int. Conf. on Spoken Language Processing, 451-454 Yokohama.

Odell, J. J. 1995. The Use of Context in Large Vocabulary Speech Recognition. Ph.D. diss., University of Cambridge.

Richardson, M., Bilmes, J., and Diorio, C. 2003. Hidden-Articulator Markov Models for Speech Recognition. *Speech Communication* 41: 511-529.

Tarsaku, P., Kanokphara, S. 2002. A Study of HMM-Based Automatic Segmentations for Thai Continuous Speech Recognition System. In Proc. Symposium on Natural Language Processing, 217-220. Thailand.